Shakespeare His Contemporaries: An Experiment in the Linguistic Annotation and Collaborative Curation of EEBO-TCP Texts

Philip R. Burns

Craig Berry

October 23, 2014

NORTHWESTERN
UNIVERSITY

Book of English

- Large, growing, collaboratively curated, and public domain corpus of written English since its earliest modern form;
- With full bibliographical detail;
- And light but consistent structural and linguistic annotation.

Early Modern English Drama

- EEBO contains transcriptions of about 800 plays written by contemporaries and near-contemporaries of Shakespeare, e.g., from before 1660.
- 631 of those TEI transcriptions manually corrected by Martin and his students as part of the Shakespeare His Contemporaries project.
- 521 received further corrections by Martin Mueller.
- 400th anniversary of Shakespeare's death in 2016: opportunity to produce "good" collection of these plays.

Shakespeare His Contemporaries

- Literary background presented by Martin Mueller in 2013 Hilda Hulme Memorial Lecture:

  http://www.youtube.com/watch?v=_1QgsRx5qHY
- Also see Martin's blog postings at:

  http://scalablereading.northwestern.edu
- Five undergraduates corrected 631 plays over period of eight weeks in summer 2013: Nayoon Ahn, Hannah Bredar, Madeline Burg, Nicole Sheriko, and Melina Yeh.

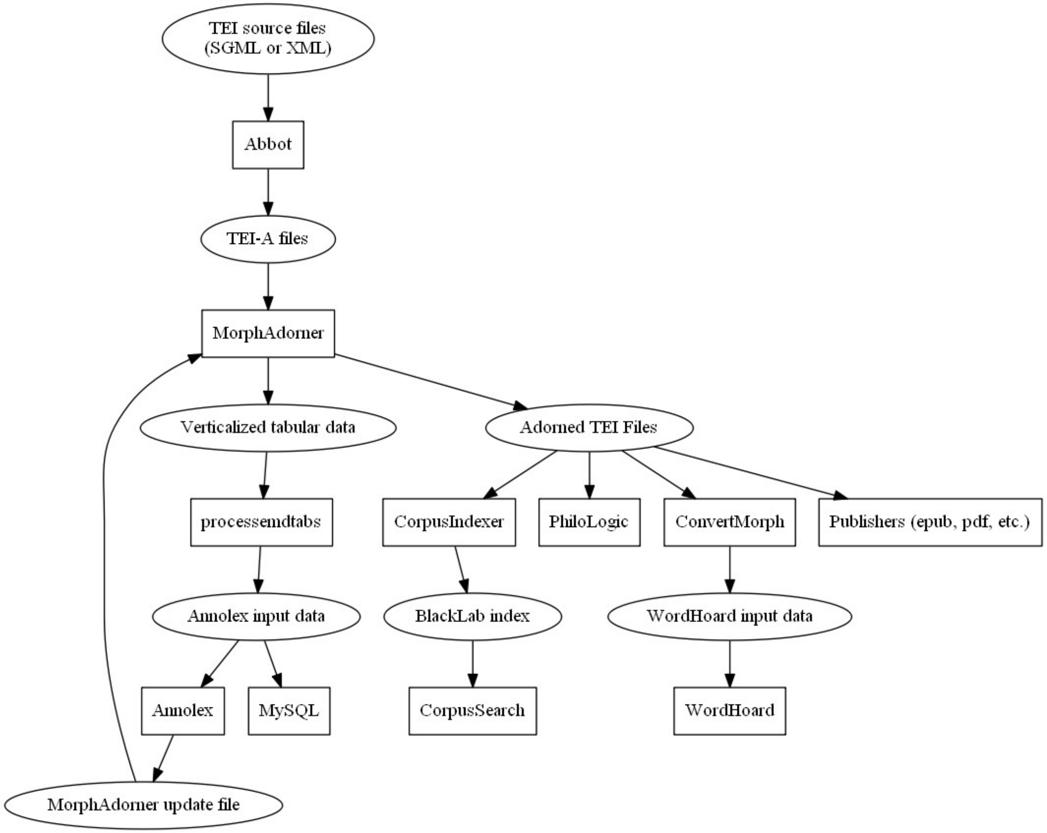
What is the SHC Experiment?

- Build environment with editorial supervision to allow for user-driven and incremental improvement of texts over time.
- Allow educated and interested readers with no special scholarly training to spot and correct many errors in an environment that provides for easy alignment of the page image with the transcribed text.

Software Used

- Abbot
- MorphAdorner
- Annolex
- MorphAdorner
- CorpusIndexer/BlackLab
- CorpusSearch
- ConvertMorph/WordHoard
- Publishers (epub, pdf, etc.)

Software Flow



SHC Processing Steps

- Convert “raw” EEBO EMD texts to TEI-A using Abbot.
- Generate adorned TEI P5 texts with lemmata, parts of speech, standard spellings using MorphAdorner.
- Generate verticalized files using MorphAdorner.
- Use verticalized files modified by processemdtabs as input to Annolex.

SHC Processing Steps (cont.)

- Use Annolex to perform token-based corrections.
- Emit Annolex corrections as tabular files.
- Merge corrections back into adorned TEI texts using MorphAdorner.
- Feed corrected adorned TEI P5 files to search engines such as Philologic and CorpusSearch (via CorpusIndexer).

SHC Processing Steps (cont.)

- Use ConvertMorph to generate WordHoard input from corrected adorned TEI P5 files.
- Generate HTML, EPUB, Kindle, etc. versions of texts from adorned files using Oxford’s TEI conversion scripts and standard etext publishing software such as kindlegen and calibre.

Abbot

- Written by Brian L. Pytlik Zillig and Stephen Ramsay in consultation with Martin Mueller.
- Normalizes TEI files to a specific TEI subset (in our case, TEI-A).
- Can also convert EEBO SGML to TEI.
- More Abbot information at http://abbot.unl.edu/cocoon/vicar/

MorphAdorner

- Terms like “annotation” and “tagging” have too many different meanings
- “Adornment” harkens back to medieval sense of adorning or illuminating a manuscripts: attaching notes and images to words, lines, paragraphs, and pages
- Morphological adornment is the process of adorning with parts of speech, lemmata, etc.

MorphAdorner Web Site

- http://morphadorner.northwestern.edu/
- Documentation in HTML, PDF, EPUB, and Kindle MOBI format
- Many online examples of MorphAdorner facilities using MorphAdorner v2 server

Basic Adornment Processes

- Tokenization
- Sentence boundary recognition
- Spelling normalization
- Part of speech assignment
- Lemmatization

If you get these correct, you can do a LOT of other things.

TCP Texts Processed

- 44,420 EEBO (Early English Books Online) texts (first 25,000 to be publicly available in 2015)
- 4,585 Evans texts (to be publicly available in June 2014)
- 2,091 ECCO texts (already publically available)
- Nebraska has the ECCO texts available in adorned and unadorned format: http://abbot.unl.edu/abbot-morphadorner/index.html

Drama Texts Processed

- EEBO texts containing more than one play split into individual plays.
- 631 plays corrected by students.
- 521 plays received further corrections by Martin Mueller during 2013/2014.

Annolex

- Simple, user-friendly web app for fixing transcription errors.
- Search features and "preselected" filter to make scum float to the top.
- Word-at-a-time view of search results based on MorphAdorner's verticalized output.
- Page images from EEBO allow reconsideration of original transcription decisions.

Annolex

- Revised transcriptions can be suggested, then later approved by the same or different user.
- Approved corrections are exported and fed back to MorphAdorner.
- Corrections and their review history comprise the "lexical annotations" that give AnnoLex its name.

Annolex

- Based on ubiquitous and developer-friendly Django, MySQL.
- Source code is open: https://code.google.com/p/annolex/

What’s Not Corrected?

Annolex is great for correcting word (token) based errors. It doesn’t currently provide for correcting other types of errors, e.g.:

- XML encoding errors.
- Bibliographic metadata problems.
- Missing pages that were never imaged.

Correction Statistics

- Before correction the SHC texts contained about 47,000 word-based errors.
- After correction by the student team, the number of errors had been reduced to about 12,500.
- Error rate reduced from 40 per 1,000 words to 10 per 1,000 words.
- Remaining errors probably need either good quality printed copy of the text or higher quality scans.

Merging Corrections

- MorphAdorner accepts corrections file from Annolex and merges the corrections back into the adorned TEI P5 files.
- MorphAdorner can provide a sequence of correction files that allow updating or downdating the TEI P5 files.

What can we do with adorned files?

- Index them for searching using Philologic, CorpusIndexer/BlackLab, Sketch engine.
- Generate “publishable” files in HTML, EPUB, PDF formats using TEI XSL scripts, calibre, kindlegen.
- Generate input for analysis programs such as WordHoard.

Corpus Search

- Philologic v4.0 processes MorphAdorned files directly.
- BlackLab library provides customizable and embeddable corpus search facilities.
- Locally written CorpusIndexer to produce BlackLab index of MorphAdorned files.
- Example BlackLab based search:

  http://devadorner.northwestern.edu/corpussearch/

Publishing to EPUB3

- EPUB3 is an HTML based publishing format popular for handheld devices (phones, tablets) as well as desktop systems.
- MorphAdorned texts can be converted to EPUB3 format using XSLT stylesheets.
- EPUB3 files can be created from the original text or the standardized text.

More Output Types

- Adorned and unadorned TEI
- Plain text, main/paratext selectable
- Verticalized tabular files and summaries
- HTML, Epub/EPub3, DocBook, MOBI, Latex, Markdown, PDF
- Sketch, TCF

Future: TEI Processing System

- TEI normalization with Abbot
- Collaborative online XML editor
- Keyword cataloguing of TCP texts that with results incorporated into the teiHeaders of documents
- Linguistic annotation with MorphAdorner
- Named entity extraction for TCP texts
- Collaborative curation with enhanced Annolex

Future: TEI Processing System

- Integrated sort and search with Philologic, BlackLab, etc.
- Online display of TEI texts with search integration.
- Integrated display of original page images if available.
- Integrated display of original illustrations.
- Generate multiple text formats – EPUB, MOBI, PDF, etc.
- Statistical methods as in WordHoard.
- RESTful API.

Questions?